# Premier League predictive learning algorithm (PLePA)

Contents

[Premier League predictive learning algorithm (PLePA) 1](#_Toc43706319)

[Glossary 2](#_Toc43706320)

[Section 1 Preparation and planning 3](#_Toc43706321)

[1.1 Project Title 3](#_Toc43706322)

[1.2 Project description 3](#_Toc43706323)

[1.2.1 What is the problem and why it is a problem 3](#_Toc43706324)

[1.2.2 Benefits of solving the problem 3](#_Toc43706325)

[1.2.3 Scope of problem 3](#_Toc43706326)

[1.2.4 Format of solution 7](#_Toc43706327)

[1.2.5 Delivery aims 7](#_Toc43706328)

[1.3 Literature review 7](#_Toc43706329)

[1.3.1 Pre TMA03 Literature 7](#_Toc43706330)

[1.4 Updated resource list 7](#_Toc43706331)

[1.5 Updated skills list 7](#_Toc43706332)

[1.6 Updated risk management 7](#_Toc43706333)

[1.7 Updated plan and reasons for change. 9](#_Toc43706334)

[1.8 LEPSI review 10](#_Toc43706335)

[Section 2 Project work to date 11](#_Toc43706336)

[2.1 Literature review 11](#_Toc43706337)

[2.2 Work done since last TMA 13](#_Toc43706338)

[Section 3 Review and reflection 13](#_Toc43706339)

[3.1 What has gone well 13](#_Toc43706340)

[3.2 What has not been so good 13](#_Toc43706341)

[3.3 Why have things not been good 13](#_Toc43706342)

[3.4 Improvement to practice 14](#_Toc43706343)

[3.5 How and why these problems and LEPSI have affected work, plan and progress 14](#_Toc43706344)

[3.6 Responses to previous tutor feedback 14](#_Toc43706345)

[References 14](#_Toc43706346)

[Bibliography 15](#_Toc43706347)

[Appendices 16](#_Toc43706348)

[Appendix 1 – Resource list 16](#_Toc43706349)

[Appendix 2 – Skill list 16](#_Toc43706350)

[Appendix 3 – Risk list 17](#_Toc43706351)

[Appendix 4 – Schedule 18](#_Toc43706352)

[Appendix 5 – Table creation script 19](#_Toc43706353)

[Appendix 6 – Main function of database reset script 20](#_Toc43706354)

[Appendix 7 – Main function of database insert script 20](#_Toc43706355)

[Appendix 8 – Research on Random Forest algorithm 20](#_Toc43706356)

[Appendix 9 – Project Journal 21](#_Toc43706357)

[Appendix 10 – Snippet of communication with tutor 22](#_Toc43706358)

|  |  |  |  |
| --- | --- | --- | --- |
| Appendix | Page | Name | Description |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Figure | Page | Name | Description |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Table | Page | Name | Description |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Glossary

PLePA – Premier League predictive learning algorithm, the name of the project.  
Premier League – The highest league in English football.  
Structured-case lifecycle – The lifecycle chosen for PLePA, it consists of multiple Conceptual Frameworks, described below.  
Conceptual Framework, also known as CF – A conceptual framework is a chunk of the lifecycle, it has a plan, process, analysis and reflection piece.   
LEPSI – Legal, Ethical, Professional and Social issues.  
Random Forest – Random Forest is one of the algorithms developed.  
K-NN – K-nearest neighbours is one of the algorithms developed.  
SDLC – Software Development Life Cycle, it is the process from beginning to end of a software project, there are many different life cycle choices.

# Section 1 Preparation and planning

## Project Title

Premier League predictive learning algorithm (PLePA)

## Project description

The project is to create a predictive algorithm for Premier League results. It will use historic data; it will use data since the 2014-15 season. It is now a trivial task to add more data if it is deemed needed because the work has been done on creating the Python scripts to insert the data and notes have been made on cleansing the data, which is a couple of simple steps. The difference between this project and similar projects out there is that all the algorithms developed will use the distance travelled for the away team as a key factor.

The result is aimed at Premier League fans and people interested in the prediction of football results. It’s also something I have an interest in solving, I do a bit of sports betting so this could prove useful for this.

The aim is to produce four separate algorithms and determine which has the most accurate results. The four algorithms I will use are Support Vector Machines (SVM), Random Forest, Naïve Bayes and K-NN. The goal is to have around a 75-80% pass rate, this will be a benefit to all football fans, fans who place bets and even possibly teams, that may be a bit far though.

If the project is unable to give good results after many iterations for each algorithm and tweaking the numbers, then it will serve as research for people who take on a similar project. There is no real issue if it not developed because it is more of a personal project and I am the only stakeholder.

There are different ways to achieve the results, the aim is to develop four separate predictive algorithms, this will give the opportunity to move on if one is not going well.

The stats data will be stored in a MySQL database and all the algorithm code will be written in Python. All code and documents will be stored in a GIT repository. A stretch goal would be to have an interface for user interaction or to pull the latest set of fixtures from a football website.

### 1.2.1 What is the problem and why it is a problem

The problem is the inability to predict football results with a good degree of accuracy. This makes things such as betting on or even just predicting premier league football results difficult. It is also interesting, for football fans and statistic fanatics, seeing the likelihood of certain teams winning in unlikely situations.

### 1.2.2 Benefits of solving the problem

The benefits of solving the problem is only that it can be used as a tool for loosely predicting premier league results. It is mainly a personal project so if it remains unsolved there is no negative impact on the wider community.

1.2.3 Scope of problem  
PLePA will have four algorithms developed and each algorithm will have a pass percentage rate, possibly with some more statistics. This will depend on what is gatherable. The algorithms will use Premier League data stored in a MySQL database. There are two stretch goals – developed a feed from a sports website to pull in latest fixtures and develop a user interface so PLePA can be distributed beyond personal use. The four algorithms will also be compared.

|  |  |  |  |
| --- | --- | --- | --- |
| Scope | Aspect of problem | Associated tasks | In or out  If out, why out? |
| Preparation stage, plan and make decisions on how the project will be carried out. | The goals give a clear indication of what the project is aiming for. This is useful for keeping the project on schedule. | Define the goals and contents of my project.  Plan the database structure to use.  Look into Python modules that will be useful.  Set up base Python project with GIT version control.  Install database software.  Find the best source for the Premier League statistics required.  Review legal, ethical, social and professional issues and decide if there are additional actions required. | In and completed. |
| Research additional material that will be useful. | Carry out research so PLePA can make informed decisions. | Research SDLC choices and decide on one for the project.  Research difference between Oracle and MySQL and decide on one of them to use.  Research similar studies for ideas and document key findings.  Research predictive algorithms. | In and completed. |
| Conceptual Framework (CF) 1 - Set up a database to store the Premier League data and the algorithm data, where required. | All algorithms will need the data to work, so it is essential work to solve the problem. | Gather required Premier League data.  Cleanse the data gathered.  Create the database.  Create a database reset script. | In and completed. |
| CF2 - Develop the Random Forest algorithm. | The first of the four algorithms to be developed for PLePA. | Research algorithm.  Plan how algorithm will be coded.  Code algorithm.  Test algorithm.  Evaluate the algorithm.  Produce statistics on algorithm. | In and completed. |
| CF3 - Develop the K-NN algorithm. | The second of the four algorithms to be developed for PLePA. | As above. | In and completed. |
| CF4 - Develop the Support Vector Machines, SVM, algorithm. | The third of the four algorithms to be developed for PLePA. | As above. | In, not yet started. |
| CF5 - Develop Naïve Bayes algorithm. | The fourth of the four algorithms to be developed for PLePA. | As above. | In, not yet started. |
| CF6 - Develop a feed from a sports website to pull in the latest fixtures. | This could allow automation of predicting upcoming results. | Research what websites could be used to pull in the Premier League data required.  Development for the data to come in from the chosen feed. | Most likely out, but not confirmed. Based on how long the first two algorithms have taken to develop, it is unlikely there will be scope to include this item because it is only a nice to have. |
| CF7 - Develop a user interface so PLePA can be distributed. | This would allow other users to use, this is not essential since it is a personal project and there are no additional stakeholders. | Sketch a user interface for screen/s to use the PLePA algorithms developed.  Create the screen/s which have been sketched. | Most likely out, but not confirmed. Based on how long the first two algorithms have taken to develop, it is unlikely there will be scope to include this item because it is only a nice to have. |

Table 1, Project scope.

1.2.4 Format of solution  
The solution will be a program which can quickly predict results for multiple games for each algorithm. Based on previous predictions, the algorithm will also be able to give a likelihood of the predictions being correct.

### 1.2.5 Delivery aims

PLePA will have four main deliverables: Premier League stats stored in a database, a working predictive algorithm to predict the results, an interface for user interaction and a feed from a football website. Each deliverable is discussed below.

* Premier League stats stored into the database must have been complete with a years’ worth of data, ideally it will have three or more years’ worth have data but a best-case scenario, time permitting, there will be five years of Premier League stats stored.
* There will be at least two predictive algorithms developed, this is a must have for PLePA to allow for comparisons between the algorithms. It should be within scope to achieve three, but, also, without any development obstacles, four is easily achievable.
* There are two nice to haves if time goes well, firstly, an interface for user interaction will be developed, there are two parts to this, the first is for the user to load up an executable, select two clubs and they will receive a prediction from the algorithm. The second would allow the user to select the algorithm to predict with.
* The second nice to have is a feed from a football website, the upcoming fixtures would be pulled from a website, the user would select the algorithm and the predicting results would be displayed.

## Literature review

### 1.3.1 Pre TMA03 Literature

Each external material used will be discussed below:

* TM351 part 1 -26 (The Open University, 2014) is the Data management and analysis module provided by the OU, this was pivotal to PLePA because it was the inspiration for the project, and it provided the skills for developing and managing a database and algorithms.
* Machine Learning Algorithms in Python (Dataflair Team, September 2018) was a useful read, it was not too old and gave a good account of developing algorithms in Python.
* Choosing a Lifecycle Model (The Open University, 2014) gave insight into multiple lifecycles that were available and discussed how to go about choosing one for the project.
* SDLC Models Explained: Agile, Waterfall, V-Shaped, Iterative, Spiral (Osetskyi, 2017) explained in detailed some of the popular SDLC choices and what they are best for.
* Structured-case: A methodological framework for building theory in information systems (J.M. Carroll et al., 2000) provided a good insight into a SDLC which was unknown to the author of PLePA, this went on to be the SDLC used for PLePA.
* Data-files: England (Football-Data,2020) is where the data used for match history stats was collected from.
* 2019-2020 Premier League Stats (FBREF, 2020) is where the stats used for league positions was collected from, they also store historical data, so the last five seasons’ worth of data was also from that website.
* The Beautiful Game: Predicting the Premier League with a random model (Nguyen, 2018) and Betting on the English Premier League (Campanelli, 2019) were research reading to see other views on solving the same problem. It was helpful to see the different approaches, what went well, and problems that were encountered.
* Legal, Social, Ethical and Professional issues (The Open University, 2012), The Data Protection Act (UK Government, 2018), Equality Act 2010 (UK Government, 2015), Computer Misuse Act 1990 (UK Government, 1990), How copyright protects your work (UK Government, n.d), What is the freedom of Information Act? (ICO, n.d) and Code of Conduct for BCS Members (BCS, 2019) were all important to read to ensure PLePA was not in breach of any Legal, Social, Ethical or Professional issues. If PLePA was in breach, these articles would have helped to resolve and steer the project in the correct way. They also helped to keep the author up to date on the government’s latest laws and guidance.
* An Implementation and Explanation of the Random Forest in Python (Koehrsen, 2018), How Random Forest Algorithm Works in Machine Learning (Synced, 2017) and How the Random Forest Works in Machine Learning (Polamuri, 2017) were all research done in preparation for development of the first algorithm, Random Forest. The articles all explained how they had developed the Random Forest algorithm in Python which is the same development language used for PLePA so this was suitable research for the project.

### 1.3.2 TMA03 Literature

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Literature | Presentation | Relevance | Objectivity | Method | Provenance | Timeliness |
| How to Implement Random Forest From Scratch in Python (Brownlee, August 2019) | Very good, each code snippet had a good explanation to what it was doing. | Very relevant, this was the algorithm being developed in the chosen scripting language. | It is the author’s views on implementing the algorithm, there will always be multiple ways to implement the algorithm. | The author has explained his method and given detailed explanation on why. | The author has lots of articles on developing algorithms in Python with good reviews. | The material was updated less than a year ago. |
| How to Implement The Decision Tree from Scratch in Python (Brownlee, December 2019) | As above. | The research carried out in How to Implement Random Forest From Scratch in Python (Brownlee, August 2019) triggered this research, it was useful to understand the development of each tree within the Random Forest. | As above. | As above. | As above. | As above. |
| Develop k-Nearest Neighbors in Python from Scratch (Brownlee, 2019) | As above. | Very relevant, this was the algorithm being developed in the chosen scripting language. | As above. | As above. | As above. | As above. |
| Implementing a Random Forest Classification Model in Python (Huneycutt, May 2018) | The presentation was clear, but the information provided was not. | Not that useful, a lot of assumed knowledge and explanations for examples not given clearly. | As above. | Not clearly explained why the author has chosen to do things in certain ways. | The author has lots of articles on Python projects. | The material was released in May 2018 with no updates, so it is not too outdated. |
| In-Depth: Decision Trees and Random Forest (VanderPlas, March 2016) | The presentation was clear, everything followed logically. | It was mostly useful, however, some of the information given was out of the scope of PLePA. | As above. | Mostly explained well, the author has assumed some knowledge. | It’s an article from the Python Data Science Handbook which is a well-known book. | It’s a few years old so some of the material may be outdated. |
| K Nearest Neighbor Algorithm In Python (Maklin, July 2019) | Very basic but a clear layout. | This was a good starting point to research K-NN, it was quite basic and didn’t go into too much depth. | As above. | There is not a lot of information, but it is the author’s own implementation. | It is posted on towardsdatascience.com so assumed to be mostly reliable. | Less than a year old so unlikely to be too outdated if at all. |
| KNN Classification using Scikit-learn (Navlani, August 2018) | Great layout, pictures and code snippets used to break up the text and help explain. | PLePA did not end up using Scikit-learn but the article was still useful for Python research on K-NN. | As above. | All information provided is explained clearly. | It is posted on DataCamp’s website so is expected to be reliable. | Under two years old so will be mostly in date still. |
| 1.6 Nearest Neighbors (Buitinck et al., n.d) | Lots of text which makes it hard to keep focus, but the occasional image did help to explain some parts. | This was mostly useful; some was outside the scope of PLePA. PLePA did not end up using the module but the information was still useful | It is the implementation of the algorithm using the specific scikit-learn Python module. | The author developed the module so will be factual on how to use it. | It is a widely used Python module so it can be trusted. | No date for this specific article but it is expected the module will be kept up to date. |

UP TO HERE

1.4 Updated resource list

The resource below, information from previous studies, has been added. This was overlooked as a resource, but it is very useful to see how other studies have investigated a similar problem. The initial resource list can be seen in Appendix 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resource | Why needed | When needed | Problems if not available | How to ensure availability |
| Information from previous studies | To understand what went well and not so well for similar projects. | Studied prior to any real task being completed, used throughout. | No knowledge from other teams’ experiences so PLePA may make similar mistakes as other studies which were not investigated. | Look on the internet and in the library early in the project. |

Table 2, Additional resources list

## 1.5 Updated skills list

There are no new skills identified for PLePA, the initial skill list can be seen in Appendix 2.

## 1.6 Updated risk management

An updated risk list is below, all other risks which have been identified previously are in Appendix 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk involved | Priority – low to high | Overcome/manage risk | What to do if not managed | Review of risk |
| Unable to gain knowledge on report writing | High | Lots of preparation and research ahead of the EMA, the TMAs are good practice too. | A poor representation of work done, won’t be reflective of the work put in in the project. | Avoid – make sure to research and ask for advice from contacts who are good at report writing. |
| Unable to obtain Premier League data | High | Ensure the data is available and download onto PC before lost. | No data available for the project, very difficult to continue | Avoid – it would be disastrous to not obtain the data required. |
| Time | Medium | Manage time correctly, try to productive when working and keep up with the schedule. | Project may fall behind, and milestones may not be met. The project could be unfinished. | Accept – manage time appropriately, try not to fall too far behind on schedule. If behind on schedule, try to catch up. |

Table 3, Additional risks list

## 1.7 Updated plan and reasons for change.

Write LEPSI review was added, this was a task that needed to be completed and needed dedicated time to it. Re-evaluate project was moved because it was not required to be done on 9th March when initially scheduled. Plan how algorithm will work for CF2 was pushed back a week due to assignment and LEPSI commitments. The initial schedule can be seen in Appendix 4.

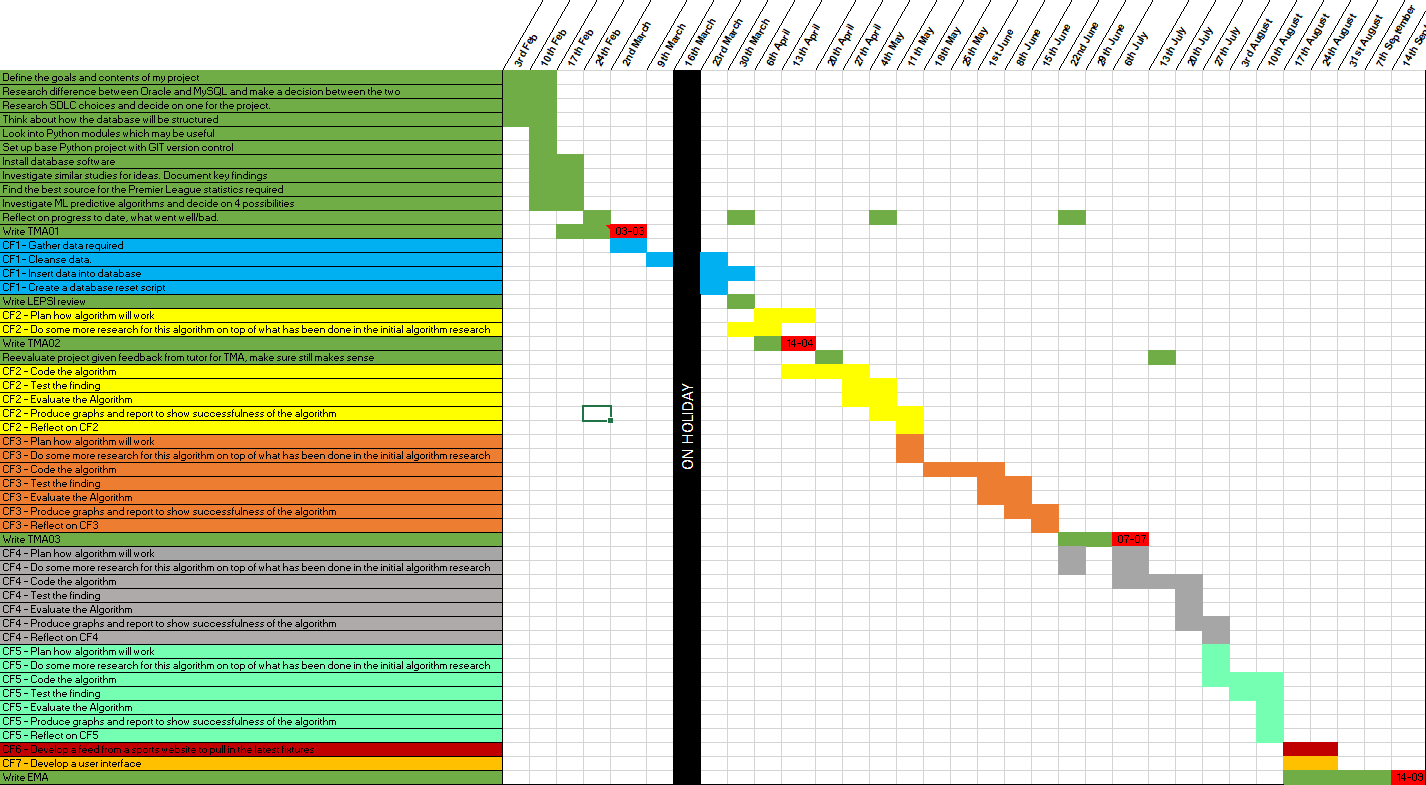


Figure 2, (Blagg, 2020) Work schedule for PLePA

## 1.8 LEPSI review

|  |  |  |  |
| --- | --- | --- | --- |
| Act / Law / Guidance | Purpose | Relevance to my project or n/a and why not applicable | How my project will be affected |
| Data Protection Act 2018 (DPA) | Control how personal information is used by organisations, businesses or the government (UK Government, 2018) | I will not be using any personal information in my project, so this is not applicable for me. | My project will be unaffected. |
| Equality Act 2010 | Protection from discrimination in the workplace and society (UK Government, 2012) | I will not be looking at people in my project so this will not be applicable for me. | My project will be unaffected. |
| Computer Misuse Act 1990 | The act was brought into place to prevent people gaining unauthorised access to computer material, commonly known as hacking. (UK Government, 1990) | My project will likely not involve connection to the internet. It may do, if I reach my stretch goal to connect to a sports website for latest fixtures. | If I reach my stretch goal for connection to the sports website to get the latest fixtures, I will ensure the connection I have is secure and authorised. It likely will be authorised since the website will be publishing it. |
| Copyright | To protect your work by preventing it being copied, redistributed, adapted and put on the internet are some of the example (UK Government, n.d) | I will need to make sure the data that I use for past football results is not protected by copyright and can be used. | I will need to ensure I am able to use the data and not breaking copyright rules. |
| Freedom of Information Act | It gives the general public access to certain information on request from the public authorities. The public authorities are also obliged to publish certain information. (ICO, n.d) | This will not affect my project because I am not working for a public authority and I do not require data from a public authority. | My project will be unaffected. |
| Protection from harm | Protect participants of studies from any harm, physical or psychological, particular care should be paid to children. (TM470 course team, 2012) | This will not affect my project because I am not using participants. | My project will be unaffected. |
| Professional codes of practice and ethics | This allows participants to understand the purpose of the study, the researchers must state their intentions. (TM470 course team, 2012) | This will not affect my project because I am not using participants. | My project will be unaffected. |
| BCS Code of Conduct | To set out standards across the board for all members of the BCS. An example is having respect for public health, privacy, security and wellbeing of others and the environment. (BCS, June 2019) | This will not affect me directly because I am not a member of the BCS, however, I should try to comply with the standards set out by the BCS such as honesty with my skillset and acknowledgment to any borrowed source code used in my project. | I must ensure that borrowed code is acknowledged, and I am not overexaggerating my skillset. |

Table 4, LEPSI review

# Section 2 Project work to date

## 2.1 Literature review

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Literature | Presentation | Relevance | Objectivity | Method | Provenance | Timeliness |
| Data-files: England (Football-Data, 2020) | Fairly easy to follow, each season’s worth of data is split between football divisions. | Very relevant, this is the exact data that is required for PLePA. | It is factual information, data from past football seasons. | The information has come from past football seasons. | Having cross checked some of the data, there are no inconsistencies. | The material was updated on 15th March, after the last game was played. |
| 2019-2020 Premier League Stats (FBREF, 2020) | Very easy to follow, the data is displayed and downloadable. | Very relevant, this is the exact data that is required for PLePA. | It is factual information, data from past football seasons. | The information has come from past football seasons. | Having cross checked some of the data, there are no inconsistencies. | The latest results are displayed so it is recent but there is no updated date. |
| The Beautiful Game: Predicting the Premier League with a random model (Nguyen, 2018) | Nice written report, broken up well with pictures and graphs. | Very relevant, it is a very similar report to what PLePA is expected to produce. | Slightly bias, only investigates a single algorithm. | Working is explained well, clear where data has come from. | It is shared on Medium and the author supports their arguments. | Fairly recent in the grand scheme of things, 2018. |
| Betting on the English Premier League (Campanelli, 2019) | Nicely written, good images to show their workings. | Very relevant, it is a very similar report to what PLePA is expected to produce. | Good comparison between different statistics and algorithms. | Working is explained, clear where data has come from. | It is shared on Medium and the author supports their arguments. | Written in the last twelve months. |
| Legal, Social, Ethical and Professional issues (The Open University, 2012) | Easy layout to understand after doing lots of OU work. | Helpful to understand LEPSI for review. | Facts about LEPSI given. | It is clear where the information has come from. | The author is reliable because it is material directly from an OU module. | It has come from the OU so should be up to date, no date given. |
| The Data Protection Act (UK Government, 2018) | Standard GOV.UK layout so easy to follow. | Relevant to understand rights to data. | Facts from the government. | Document from the government so it is correct. | Directly from the government so very reliable. | Expected to be fully up to date on a government website. |
| Equality Act 2010 (UK Government, 2015) | Standard GOV.UK layout so easy to follow. | Relevant to understand the Equality Act but probably not relevant to PLePA | Facts from the government. | Document from the government so it is correct. | Directly from the government so very reliable. | Expected to be fully up to date on a government website. |
| Computer Misuse Act 1990 (UK Government, 1990) | Standard GOV.UK layout so easy to follow. | Relevant to understand the Computer Misuse Act, probably not relevant to PLePA | Facts from the government. | Document from the government so it is correct. | Directly from the government so very reliable. | Expected to be fully up to date on a government website. |
| How copyright protects your work (UK Government, n.d) | Standard GOV.UK layout so easy to follow. | Relevant to understand the copyright protection for data that is used. | Facts from the government. | Document from the government so it is correct. | Directly from the government so very reliable. | Expected to be fully up to date on a government website. |
| What is the freedom of Information Act? (ICO, n.d) | Ok, not the easiest to follow through | Relevant to understand freedom of information act but PLePA will not be using personal data. | Facts from public body which report directly to parliament. | Public body reporting to parliament so expected to be correct. | Directly from public body which report directly to parliament. | No date but expected to be kept up to date since it is from the ICO. |
| Code of Conduct for BCS Members (BCS, 2019) | Not very easy to follow. | Important to try and comply with BCS standard so useful to know their standards. | BCS create their rules so their website is the best place to obtain the information. | Following BCS rules so expected their website will be correct. | Directly from BCS website. | Date found is 2019 but expected to be up to date. |
| An Implementation and Explanation of the Random Forest in Python (Koehrsen, 2018) | Easy layout to follow, broken up with graphs and images. | Helpful to understand Random Forest and how it can be implemented, particularly in Python. | It is an explanation of Random Forest, so it is objective. | Information is obtained using author’s work and explanation is clear. | It is a post shared on Medium so expected to be correct. It is not being considered alone. | Under two years old so not really outdated. |
| How Random Forest Algorithm Works in Machine Learning (Synced, 2017) | Nice layout with text broken up with graphs and images. | Relevant for Random Forest, however, not too much was gained from reading this. | It is an explanation of Random Forest, so it is objective. | Information is building on the work of Polamuri (Polamuri, 2017). | It is a post shared on Medium so expected to be correct. It is not being considered alone. | Just under three years old, some new knowledge may have been built on this. |
| How the Random Forest Works in Machine Learning (Polamuri, 2017) | Very good layout and information is clearly explained. | Very useful to help understand Random Forest and Decision Trees. | Explanation of Random Forest and Decision Trees | Very clearly explained how information is collected. | It is a post shared on Medium so expected to be correct. It is not being considered alone. | Just under three years old, some new knowledge may have been built on this. |

Table 5, Literature review

## 2.2 Work done since last TMA

All the data to be used in PLePA has now been gathered (FBREF, 2020). The decision was to download all data since the season that ran from 2014 to 2015. Once the data had been gathered, it need to be cleansed ready to be inserted into the database. Upon investigating the data, it did not look like any data cleansing was required.

The stadium data was also collected manually by looking at the location on Google Maps and finding the coordinates.

The database table creation was then created, as seen in Appendix 5. This was a straightforward task because the structure of the database had already been decided on.

Once the tables were created and all data was collected, work begun on creating the Python script to insert the data into the database. This revealed an issue in the data, the results data contained Norwich, but the historic season results contained Norwich City. This caused some issues in the Python script when inserting because it was looking for a name which did not exist in the database. This involved some cleansing of the data which was overlooked initially. There were some other small issues with the data which triggered the need for a database reset script. A database reset script was created which was able to quickly reset the database to a state with the tables created but no data yet inserted, the main function of the database reset script can be seen in Appendix 6. The main function of the completed database insert script can be seen in Appendix 7.

After doing the above tasks, this meant I had completed the evaluation criteria to have 5 years, nice to have, worth of data inserted into the database.

The last chunk of work that was done, seen in Appendix 8, was research into the Random Forest algorithm. This has been decided to be the first algorithm to develop in PLePA. The research can be seen in Appendix 8, it involved looking into multiple sources ((Koehrsen, 2018), (Synced, 2017), (Polamuri, 2017)).

The day to day breakdown of work done since the last assignment can be seen in the project journal in Appendix 9.

# Section 3 Review and reflection

## 3.1 What has gone well

Following the schedule has worked very well, this has kept the project on track and ensured that tasks are being completed when they should be done. It is hard to comment on the usefulness of the Software Development Life Cycle (SDLC) chosen so far because it has not really played a part. The first CF is now completed, inserting the data into the database. Unfortunately, the house move has been cancelled because of COVID-19, however, for the project, this means that there is now more time available at the moment. The house move may proceed once, if, COVID-19 subsides. The coding part of the project is now starting with the database scripts, these tasks have been enjoyable to do.

## 3.2 What has not been so good

The oversight on the data cleansing, where Norwich and Norwich City were used in different data sets. Setting time to work and sometime procrastinating for too long has caused some wasted time.

## 3.3 Why have things not been good

The oversight caused a small set back as it was not initially obvious what the problem was, I had to add some logging into my Python code to figure out where it was failing.

At the moment, there seems to be more time available to do the project, however, it is not always easy to focus with everything that is currently going on in the world. Initially, I thought I would be able to be much more productive, however, just because more time is available it does not necessarily mean you can work on the project during those hours because a lot of time is spent procrastinating.

## 3.4 Improvement to practice

I have decided, based on the procrastination issue, to do work when I am in the mood rather than setting myself specific time to do work. Obviously, if this then means the project starts to fall behind then I will need to set some specific time to do the project, however, I do not think this will be an issue.

## 3.5 How and why these problems and LEPSI have affected work, plan and progress

Due to the nature of PLePA, there have no been real issues with LEPSI. The only issue will be to copyright this work from other people if necessary.

There has been no real setback from the issues described in section 3.2, there have just been plans put in place to improve these in the future.

Overall, the project is going very well and according to schedule.

## 3.6 Responses to previous tutor feedback

A snippet from the communication with my tutor can be seen in Appendix 10, I used the feedback from Ju to improve my LEPSI review. There has been ongoing communication with Ju throughout the module, I have always taken on board and adapted my work based on Ju’s feedback.

I have included table and figure names now, based on feedback from Ju in TMA01. I have also tried to include better referencing and included all literature since TMA01 in the literature review.

**[4031 Words]**

# References

UK Government (2018) *The Data Protection Act* [Online]. Available at <https://www.gov.uk/data-protection> (Accessed 1st April)

UK Government (2015) *Equality Act 2010: guidance* [Online]. Available at <https://www.gov.uk/guidance/equality-act-2010-guidance> (Accessed 1st April)

UK Government (1990) *Computer Misuse Act 1990* [Online]. Available at <http://www.legislation.gov.uk/ukpga/1990/18> (Accessed 1st April)

UK Government (n.d) *How copyright protects your works* [Online] Available at <https://www.gov.uk/copyright> (Accessed 1st April)

Information Commissioner’s Office (ICO) (n.d) *What is the Freedom of Information Act?* [Online]. Available at <https://ico.org.uk/for-organisations/guide-to-freedom-of-information/what-is-the-foi-act/> (Accessed 1st April)

TM470 course team (2012) *‘Legal, Social, Ethical and Professional issues’ in TM470 study material, The Open University, Milton Keynes* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1184585> (Accessed 1st April)

BCS, The Charted Institute for IT (June 2019) *Code of Conduct for BCS Members* [Online] Available at <https://cdn.bcs.org/bcs-org-media/2211/bcs-code-of-conduct.pdf> (Accessed 1st April)

FBREF (February 2020) *2019-2020 Premier League Stats* [Online]. Available at <https://fbref.com/en/comps/9/Premier-League-Stats> (Accessed 24th February)

# Bibliography

Dataflair Team (September 2018) *8 Machine Learning Algorithms in Python* [Online]. Available at <https://data-flair.training/blogs/machine-learning-algorithms-in-python/> (Accessed 16th February 2020)

J.M.Carroll, P.A.Swatman (2000) *Structured Case: A Methodological Framework for Building Theory in Information Systems Research*. Available at <https://pdfs.semanticscholar.org/caff/950b0efc8f23e17e7de171d8836c8dfa1044.pdf>(Accessed 17th February 2020)

The Open University (2014) *Choosing a Lifecyle Model*, *TM470* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1557365>(Accessed 2nd February 2020)

Victor Osetskyi (2017) *SDLC Models Explained: Agile, Waterfall, V-Shaped, Iterative, Spiral* [Online]. Available at <https://medium.com/existek/sdlc-models-explained-agile-waterfall-v-shaped-iterative-spiral-e3f012f390c5>(Accessed 14th February 2020)

The Open University (2019) ‘Data mining I: classification tasks’ *The k-nearest neighbours algorithm, TMA351* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1349974&section=3> (Accessed 24th February 2014)

The Open University (2019) ‘Data mining II: clustering tasks’ *Partitional clustering: the k-means algorithm, TM351* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1349975&section=2.2> (Accessed 24th February 2014)

Football-Data (February 2020) *Data-files: England* [Online]. Available at <http://www.football-data.co.uk/englandm.php> (Accessed 24th February)

FBREF (February 2020) *2019-2020 Premier League Stats* [Online]. Available at <https://fbref.com/en/comps/9/Premier-League-Stats> (Accessed 24th February)

Tuan Doan Nguyen (August 2018) *The Beautiful Game: Predicting the Premier League with a random model* [Online]. Available at <https://towardsdatascience.com/o-jogo-bonito-predicting-the-premier-league-with-a-random-model-1b02fa3a7e5a> (Accessed 24th February)

Nick Campanelli (May 2019) *Betting on the English Premier League* [Online]. Available at <https://towardsdatascience.com/betting-on-the-english-premier-league-making-money-with-machine-learning-fb6938760c64> (Accessed 24th February)

TM470 course team (2012) *‘Legal, Social, Ethical and Professional issues’ in TM470 study material, The Open University, Milton Keynes* [Online]. Available at <https://learn2.open.ac.uk/mod/oucontent/view.php?id=1184585> (Accessed 1st April)

UK Government (2018) *The Data Protection Act* [Online]. Available at <https://www.gov.uk/data-protection> (Accessed 1st April)

UK Government (2015) *Equality Act 2010: guidance* [Online]. Available at <https://www.gov.uk/guidance/equality-act-2010-guidance> (Accessed 1st April)

UK Government (1990) *Computer Misuse Act 1990* [Online]. Available at <http://www.legislation.gov.uk/ukpga/1990/18> (Accessed 1st April)

UK Government (n.d) *How copyright protects your works* [Online]. Available at <https://www.gov.uk/copyright> (Accessed 1st April)

Information Commissioner’s Office (ICO) (n.d) *What is the Freedom of Information Act?* [Online]. Available at <https://ico.org.uk/for-organisations/guide-to-freedom-of-information/what-is-the-foi-act/> (Accessed 1st April)

BCS, The Charted Institute for IT (June 2019) *Code of Conduct for BCS Members* [Online] Available at <https://cdn.bcs.org/bcs-org-media/2211/bcs-code-of-conduct.pdf> (Accessed 1st April)

Will Koehrsen (August 2018) *An Implementation and Explanation of the Random Forest in Python* [Online] Available at <https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76> (Accessed 5th April)

Synced (October 2017) *How Random Forest Algorithm Works in Machine Learning* [Online] Available at <https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning-3c0fe15b6674> (Accessed 5th April)

Saimadhu Polamuri (May 2017) *How the Random Forest Algorithm Works in Machine Learning* [Online] Available at <https://dataaspirant.com/2017/05/22/random-forest-algorithm-machine-learing/> (Accessed 5th April)

# Appendices

## Appendix 1 – Resource list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Resource | Why needed | When needed | Problems if not available | How to ensure availability |
| Time | Complete any task for the project. | Throughout project. | The less time available, the fewer tasks can be completed. | Cannot ensure time but can try to get ahead on tasks in case of any complications. |
| PC | To complete all tasks, requires pc availability. | Throughout project. | Whilst no PC is available, most tasks will be unable to be achieved. | I have a desktop and laptop. As a last resort, I can also use my work laptop. |
| Python | Write the code for the algorithms. | During CF2-5. | The code for the algorithms will not be able to be written. | It is installed on all PCs available. |
| MySQL | Store the data used for algorithms. | Set up in CF1 but will be required from CF1 – CF7. | The algorithms will have no database to read from. | It is installed on all PCs available. |
| Premier League data | The project is using data from the Premier League for the algorithms. | Pre CF1, it will then be stored in the MySQL database. | No data to be used for the algorithms. | Once gathered and stored in the database, store a copy of the database. |

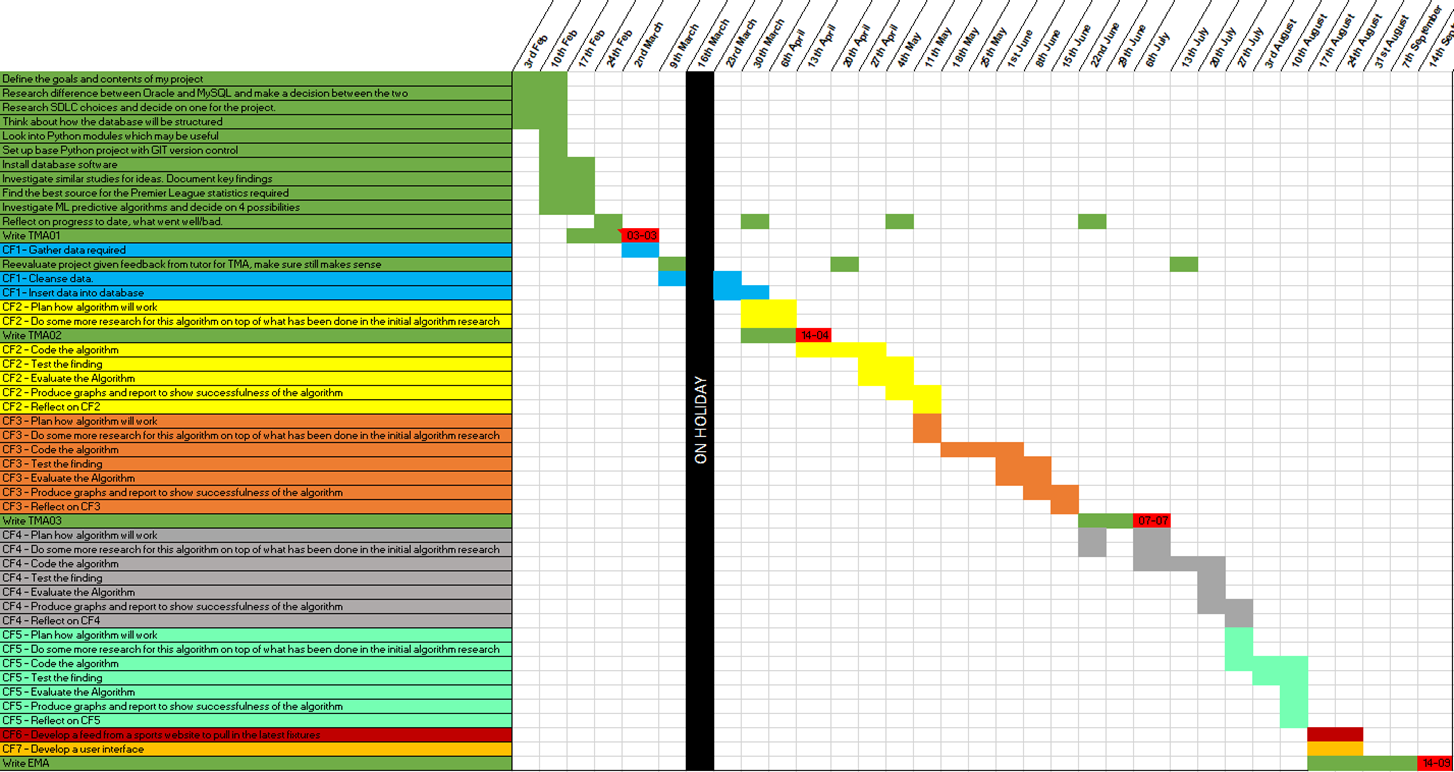
## Appendix 2 – Skill list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Skill | Why needed | When needed | Problems if not available | How to ensure availability |
| Python coding | Write the code for the algorithms. | During CF2-5. | The code for the algorithms will not be able to be written. | Research unused additional Python modules which are required. |
| SQL coding | Write the code for the database. | During CF1. | Unable to create the database structure and insert the data. | Used in everyday work, fluent in SQL. |
| Time management | Ensure milestones are met. | Throughout project. | Milestones may be missed; project will be incomplete. | Research good time management ideas and seek advice from people with good time management. |
| Report writing | To write TMA1-3 and EMA. | For all assignments but mainly the EMA. | The quality of writing for TMAs and EMAs will be lacking. | Research and check previous module’s advice on report writing and read course booklet on the subject |
| Researching | For several tasks to investigate best approach | Throughout project. | Project will suffer due to decisions being made without right information. | Look into ideas for researching, the best approaches to research and read course booklet on the subject |

## Appendix 3 – Risk list

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk involved | Priority – low to high | Overcome/manage risk | What to do if not managed | Review of risk |
| Planning on a house move. | Medium | Get ahead on project when possible, manage time during move. | Make up the time after the house move is complete. | Accept – try to manage time accordingly around busy times for house move. |
| Work travel, 2-4 days a month. | Low | Manage workload accordingly around work travel to get ahead. | Try to find a bit of extra time following the travel to catch up on project. | Accept – try to get ahead before work travel and catch up if needed afterwards. |
| Desktop and laptop failure | Low | Data stored at GitHub and locally on both machines. | Library work or purchase new laptop/desktop to replace. | Avoid – Unlikely that both desktop and laptop will fail. |
| Knowledge gap on new Python modules | High | Research the modules required, also have a Python expert at my work who I can consult with. | Lower the detail of the algorithms used. | Accept – start researching required modules early and learn how to use. |

## Appendix 4 – Schedule



## Appendix 5 – Table creation script

CREATE TABLE stadium (

stadium\_id\_pk INT AUTO\_INCREMENT PRIMARY KEY,

stadium\_name VARCHAR(255) NOT NULL,

x\_coord DECIMAL(13,8) NOT NULL,

y\_coord DECIMAL(13,8) NOT NULL

);

CREATE TABLE team (

team\_id\_pk INT AUTO\_INCREMENT PRIMARY KEY,

name\_pk VARCHAR(255) NOT NULL,

stadium\_id\_fk INT,

CONSTRAINT fk\_stadium

FOREIGN KEY (stadium\_id\_fk)

REFERENCES stadium(stadium\_id\_pk)

);

CREATE TABLE season\_overview (

team\_id\_pk\_fk INT,

season\_pk INT,

position INT,

goals\_for INT,

goals\_against INT,

wins INT,

draws INT,

losses INT,

PRIMARY KEY (team\_id\_pk\_fk, season\_pk),

CONSTRAINT fk\_team\_id

FOREIGN KEY (team\_id\_pk\_fk)

REFERENCES team(team\_id\_pk)

);

CREATE TABLE game (

home\_team\_id\_pk\_fk INT,

away\_team\_id\_pk\_fk INT,

season\_pk INT,

home\_team\_score INT,

away\_team\_score INT,

PRIMARY KEY (home\_team\_id\_pk\_fk, away\_team\_id\_pk\_fk, season\_pk),

CONSTRAINT fk\_home\_team\_id

FOREIGN KEY (home\_team\_id\_pk\_fk)

REFERENCES team(team\_id\_pk),

CONSTRAINT fk\_away\_team\_id

FOREIGN KEY (away\_team\_id\_pk\_fk)

REFERENCES team(team\_id\_pk)

);

## Appendix 6 – Main function of database reset script

def main():  
 plepa\_db = mysql.connector.connect(  
 host='localhost',  
 user='root',  
 passwd='PLePApw',  
 database='plepa'  
 )  
  
 cursor = plepa\_db.cursor()  
  
 drop\_tables(cursor, plepa\_db)  
 create\_tables(cursor, plepa\_db)  
  
 plepa\_db.close()

## Appendix 7 – Main function of database insert script

def main():  
 plepa\_db = mysql.connector.connect(  
 host='localhost',  
 user='root',  
 passwd='PLePApw',  
 database='plepa'  
 )  
  
 cursor = plepa\_db.cursor()  
  
 insert\_data('stadium\_team', STADIUM\_DATA, cursor, plepa\_db)  
 insert\_data('game', GAME\_DATA, cursor, plepa\_db)  
 insert\_data('season\_overview', SEASON\_OVERVIEW\_DATA, cursor, plepa\_db)  
  
 plepa\_db.close()

## Appendix 8 – Research on Random Forest algorithm

Random Forest algorithm

Decision tree

* Decision tree is a non-linear model built using linear boundaries.
* Completely learning data can be a downside because it can lead to overfitting.
  + Overfitting occurs when the model is very flexible (high capacity).
  + A flexible model has high variance because the learned parameters vary considerably from the training data.
* Gini Impurity represent the probability that a randomly selected node will be incorrectly classified.

Random forest

* Two key concepts:
  + Random sampling of training data points when building trees.
  + Random subset of features considered when splitting nodes.
* Each tree goes through testing and could use some samples more than once.
* Overall, the entire forest will have lower variance (good) without increasing bias (bad).
* Bias-variance trade-off is a common issue with machine learning, it is the balance between giving a model high flexibility (high variance) that can learn training data but cannot generalize new data vs method which cannot learn the training data (high bias). A random forest reduces variance by use multiple trees.

## Appendix 9 – Project Journal

-------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Break in work due to being busy with beginning house move – selling house/seeing houses**

-------------------------------------------------------------------------------------------------------------------------------------------------------------------

**7th Mar**

Tutorial in the morning.  
Gathered data required for PLePA.  
Update journal for 7th Mar.

**9th Mar**

Created the script for the database table creation.  
Update journal for 9th Mar.

-------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Break in work due to being on holiday, arrived home early due to Covid-19.   
House purchase fell through due to buyers of our property pulling out due to Covid-19**

-------------------------------------------------------------------------------------------------------------------------------------------------------------------

**19th Mar**

Checked over data, no cleansing required. All data in format ready to be inserted into database.  
Oversight on stadium data, I collected and stored this.  
Created database inserts for the stadium data in a Python script.  
Update journal for 19th Mar.

**23rd Mar**

More work on database inserts for stadium data and game data, hit a bug which needs to be fixed.  
Update journal for 23rd Mar.

**24th Mar**

Finished insert for game data, fixed bug where there were inconsistencies with names that I did not spot before such as Norwich in game data but Norwich City in stadium data.  
 **25th Mar**

Cleaned season overview data and wrote database inserts for the data.  
Update journal for 24th and 25th Mar

**28th Mar**

Added notes for TMA02 from tutorial attended.  
Update journal for 28th Mar

**1st Apr**

Started work on LEPSI review.  
Update journal for 1st Apr

**2nd Apr**

Sent LEPSI to tutor for review.  
Update journal for 2nd Apr

**4th Apr**

Completed LEPSI work based on feedback from tutor.  
Update journal for 4th Apr

**5th Apr**

Research into first algorithm, Random Forest.

**8th Apr**

Work on TMA02

**11th Apr**

Work on TMA02

**12th Apr**

Work on TMA02  
Updated journal for 5th, 8th, 11th and 12th April.

## Appendix 10 – Snippet of communication with tutor

Hi Ju,

My progress update.

Progress made - Spoke a bit about this in the tutorial. I'm now at a good point to start writing the algorithms. I created the database insert Python scripts. This involved a bit of work and resetting of the database a few times so I also created a database reset scripts that deletes the tables and then recreates them. This means I can always just run two scripts to get back to the starting point if the data got screwed up. I noticed some data mismatches even though I got my data from the same place. As an example, in some files, there was Norwich City and some it was just Norwich. I have also done LEPSI review and got a draft to provide to you.

Decisions taken - As above, I created a database reset script. I haven't made any other changes that have taken me off course.

What has gone well - The Python scripts have gone well, I enjoyed writing them and they are working very well.

What has failed to go well - Oversights on the data mismatch caused me some issues because I did not spot the issue straight away. The error message was not clear.

What you aim to do next - I am now going to start doing more research on top of what I've already done for my first algorithm. After this, I will begin development. However, I think after I've done the research I will work on TMA02 to get that done rather than stop and starting development.

Questions - Review of my LEPSI please.

Kind regards,

Rob

---------------------------------------------------------------------------------------------------------------------------------------------------

I hope the Canaries' data were good! Probably not!

the Lespi review is good

one thing I might add is that the BCS expect you to be honest about your skillset and to acknowledge code borrowed from outside sources so you could expand on that

Cheers

Ju